2

A PRELIMINARY TEST FOR STRUCTURE IN LARGE, HIGH-DIMENSIONAL DATA SETS

BY

FRED W. HUFFER and CHEOLYONG PARK

TECHNICAL REPORT NO. 447
SEPTEMBER 5, 1991

PREPARED UNDER CONTRACT
N00014-89-J-1627 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A PRELIMINARY TEST FOR STRUCTURE IN LARGE, HIGH-DIMENSIONAL DATA SETS

BY

FRED W. HUFFER and CHEOLYONG PARK

TECHNICAL REPORT NO. 447
SEPTEMBER 5, 1991

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A Preliminary Test for Structure in Large, High-Dimensional Data Sets

Fred W. Huffer and Cheolyong Park

1 Introduction

In this report we suggest a general approach and a specific test statistic which may help to detect the presence of 'interesting' multivariate structure in a large, high-dimensional data set. Our goal is to supply a preliminary test which is fairly easy to compute and which might predict whether it is worthwhile to carry out more computationally intensive procedures such as projection pursuit or cluster analysis. The basic idea underlying this approach is that a data set (or distribution) in which the coordinates (covariates) are independent is 'boring'. In such a data set, the multivariate structure is entirely determined by the marginal distributions. For example, if a data set with independent coordinates contains clusters, these clusters can be detected by merely examining the marginal distributions. More generally, a data set may be considered 'boring' if there is some simple transformation which converts it into a data set with independent coordinates. For example, an appropriate linear transformation will convert the multivariate normal (MVN) distribution into a distribution with independent normal coordinates, and thus the MVN distribution is 'boring'. If a distribution is 'boring' in the sense we have outlined, the multivariate structure is trivial and there is no point in carrying out procedures such as projection pursuit or cluster analysis.

Our approach, put briefly, is as follows: Given a data set, we attempt to find a simple transformation which converts it into a new data set in which the coordinates are (at least approximately) independent. After transforming the data, we test the null hypothesis of independence by discretizing each coordinate and analyzing the resulting categorical data as a contingency table. We can compare the cell counts in this contingency table with those expected under independence and, if a formal test statistic is desired, we can employ the usual chi-squared test of independence for contingency tables. If we resoundingly reject the hypothesis of independence, then the data set has 'interesting' multivariate structure and more computationally intensive procedures should then be used to determine the form of this structure. If we fail to reject the hypothesis of independence (and there is also no evidence of structure in the usual bivariate scatter plots of the data), then the data set is probably 'boring' and further exploration with expensive techniques might not be worthwhile.

There may be very few high-dimensional *raw* data sets which are 'boring' in the sense given above. However, the same approach can also be used to examine residuals obtained after model fitting. Typically, if the model is correctly chosen, one expects the residuals to be without structure. The methods we present may prove useful in detecting any remaining structure in the residuals.

Any preliminary examination of a data set should include studying the marginal distributions and looking at numerous bivariate scatter plots. The procedure described in this report does not in any way replace these elementary techniques.

2 Detailed Description of the Method

Let $\mathbf{X}$ be an $n \times p$ data matrix with elements x_{ij}. The n rows of $\mathbf{X}$ are obtained by random sampling from some p-variate population having a continuous joint distribution.

A. **Transform the Data:** The data should be transformed to remove known, obvious or suspected dependence/structure. (Our primary goal is to discover the existence of surprising or unsuspected multivariate structure, hence we typically try to remove known or obvious dependence such as that indicated in the correlation matrix or in bivariate scatter plots.) There are an infinite variety of transformations which can be used; we shall discuss a few of the possibilities later. In the remainder of this description, the transformed data will be denoted by $\mathbf{Z}$, an $n \times p$ matrix with elements z_{ij}.

B. **Discretize the Data:** Choose an integer d. Replace the continuous-valued quantities z_{ij} by discrete-valued quantities t_{ij} which take on the values $1, 2, \ldots, d$. This is accomplished by dividing the values in each column of $\mathbf{Z}$ into d groups of equal size n/d, that is, we set $t_{ij} = k$ if $(k-1)(n/d) < r_{ij} \leq k(n/d)$ where r_{ij} is the rank of z_{ij} among the values of the j^{th} column of $\mathbf{Z}$. The matrix with entries t_{ij} will be called $\mathbf{T}$. To avoid complications, we shall always assume that d divides n exactly. In practice, if we are given a data set not exactly divisible by d, we simply throw out a few observations (at most $d-1$) chosen at random. Since d is small (typically $2 \leq d \leq 4$), we lose little by doing this.

C. **Form a Contingency Table:** There are d^p possible p-vectors $\pi = (\pi_1, \pi_2, \ldots, \pi_p)$ with $1 \leq \pi_i \leq d$ for all i. These vectors may be regarded as cells in a $d \times d \times \cdots \times d$ contingency table. For each cell π we compute the cell count U_π which we define to be the number of observations (rows of $\mathbf{T}$) which 'belong' to π. More formally, $U_\pi = \#\{i : t_{i\cdot} = \pi\}$ where $t_{i\cdot}$ denotes the i^{th} row of $\mathbf{T}$.

D. **Study the Distribution of Cell Counts:** Now we study the frequency distribution of the d^p cell count values U_π and compare this with the frequency distribution expected under the null hypothesis of independence. Let M_k denote the number of cells containing exactly k observations, that is, $M_k =$

$\#\{\pi : U_\pi = k\}$. It is clear that

$$\overline{U} \equiv \frac{1}{d^p} \sum_\pi U_\pi = \frac{n}{d^p} \tag{2.1}$$

The values of M_k which are expected under independence can usually be well approximated by a Poisson frequency distribution with a mean of n/d^p;

$$M_k \approx d^p \left(\frac{\lambda^k}{k!} e^{-\lambda} \right) \text{ with } \lambda = \frac{n}{d^p}. \tag{2.2}$$

If the observed values of M_k differ greatly from the expected Poisson frequencies, this is evidence for the existence of 'interesting' higher dimensional structure.

It is intuitively clear that the presence of 'interesting' structure in the data will tend to increase the variability of the cell counts U_π. This suggests using the sample variance

$$W \equiv \frac{1}{d^p} \sum_\pi \left(U_\pi - \frac{n}{d^p} \right)^2 \tag{2.3}$$

of the cell counts as a test statistic for the existence of structure. This statistic is proportional to the usual chi-squared statistic for testing independence in contingency tables. We reject the hypothesis of independence when W is sufficiently large. We have obtained approximations μ_w and σ_w^2 (given later) for the mean and variance of W under the null hypothesis. These may be used to conduct rough hypothesis tests based on

$$z = \frac{W - \mu_w}{\sigma_w}. \tag{2.4}$$

These tests must be used with caution; with the usual α-levels of .05 or .01 they are likely to detect differences from the null hypothesis which are too small to be of any practical significance. This is typical of statistical testing in any situation involving large sample sizes. Rather than testing, it may be more useful to directly compare the magnitudes of W and μ_w, perhaps in terms of the ratio W/μ_w.

Transformation T1

Step A is obviously the only difficult step. One possible transformation is to replace the raw data $\mathbf{X}$ by the principal components $\mathbf{Z} = \mathbf{X\Gamma}$. Here $\mathbf{\Gamma}$ is a $p \times p$ orthogonal matrix which diagonalizes the sample covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$; $\mathbf{\Gamma^t \Sigma \Gamma}$ is a diagonal matrix. In the Examples section, we shall refer to this transformation as T1. This sort of transformation removes the correlation structure in the data and is a common initial step in many statistical procedures; see, for example, Friedman (1987). In our situation, the rationale for the transformation is as follows. Suppose there exists an orthogonal coordinate system in which the coordinates are actually independent, that is, an observation $X = (X_1, X_2, \ldots, X_p)$ is obtained as $X = Y\Lambda^t$ where Λ is an orthogonal matrix and the coordinates $(Y_1, Y_2, \ldots, Y_p)$ of Y are independent. If $\text{Var}(Y_1) > \text{Var}(Y_2) > \cdots > \text{Var}(Y_p)$ and the sample size n is sufficiently large, then $\mathbf{\Gamma} \approx \mathbf{\Lambda}$ and the principal components $\mathbf{Z}$ will be approximately equal to the independent coordinates $\mathbf{Y}$. However, if $\text{Var}(Y_i) = \text{Var}(Y_j)$ for some $i \neq j$ (or even if there is approximate equality), then the principal components $\mathbf{Z}$ will not usually be very close to the values $\mathbf{Y}$. So this transformation does not always succeed.

Transformation T2

Another transformation which may be useful is to separately transform each of the marginal distributions (columns of $\mathbf{X}$) to normality, and then further transform the data (using a linear transformation) so that the variables are uncorrelated. Stating this in detail we have:

1. Transform each variable (column) to normality. This is done in the usual way. Let r_{ij} denote the rank of x_{ij} among the values of the j^{th} column (variable). Replace x_{ij} by its normal score $y_{ij} = \Phi^{-1}((r_{ij} - \frac{1}{2})/n)$ where Φ is the standard normal distribution function. The $n \times p$ matrix of normal scores y_{ij} will be denoted by $\mathbf{Y}$. (This use of $\mathbf{Y}$ is unrelated to the earlier usage.)

2. Now apply a linear transformation to $\mathbf{Y}$ chosen so that the new variables $\mathbf{Z}$ are uncorrelated and have variance equal to one. That is, choose a $p \times p$ matrix $\mathbf{A}$ such that the transformed data $\mathbf{Z} = \mathbf{YA}$ has a sample covariance matrix equal to the identity matrix $\mathbf{I}$. There are many correct choices for the matrix $\mathbf{A}$.

In the examples section, we refer to this transformation as T2. To motivate this procedure, suppose that we have sampled $\mathbf{X}$ from a population which is roughly similar to the multivariate normal (MVN) distribution (and therefore uninteresting). In this case, there is reason to hope that the initial transformation to marginal normality will cause the data to closely resemble a sample from an MVN population. Then the linear transformation to remove the correlation will convert the data into what is essentially a sample from the MVN distribution with covariance matrix equal to the identity, and for this distribution the coordinates are independent as desired.

The previous two transformations are of general utility. However, in some cases one may have to tailor the transformation to the particular data set. For example, if a bivariate scatter plot of variable i (X_i) versus variable j (X_j) reveals a curved, approximately quadratic relationship between X_i and X_j, then a reasonable transformation might replace X_i by the residuals from the regression of X_i on X_j and X_j^2. This transformation would remove the curvature from the scatter plot.

3 Examples

Sampling from the Multivariate-Normal Distribution

The MVN distribution is surely the most boring distribution. We consider it boring because a linear transformation converts it into a distribution with independent coordinates. Researchers studying projection pursuit methods consider it boring for different reasons (see Huber (1985)). We shall use the MVN distribution to study the null behavior of our methods.

In our first example, $\mathbf{X}$ is a 1024×10 matrix composed of independent columns generated from a standard normal distribution. This data has no structure of any sort. If we suspected this in advance, we would skip the transformation (Step A) and just carry out the remaining steps B to D. We have a simple computer program which carries out steps B to D. In this case we obtain the output:

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
               0      1      2     3    4    5    6    7    8
Observed 377.00 387.00 169.00 73.00 11.0 6.00 1.00 0.00 0.00
Expected 376.71 376.71 188.35 62.78 15.7 3.14 0.52 0.07 0.01

The moments of the distribution of cell counts are:
           mean variance skewness kurtosis
Observed    1   1.0332  1.09910  1.35996
Expected    1   0.9893  0.98396  0.94713

The z-score for the variance of cell counts =  1.004
```

We have chosen to set $d = 2$; the computer output reports this as the 'number of cuts'. This means we have divided the data space into $d^p = 2^{10} = 1024$ cells. There are also $n = 1024$ observations (rows of $\mathbf{X}$), so that the average number of observations per cell is $n/d^p = 1$. The output lists the 'observed' frequency distribution: 377 cells are empty, 387 cells contain exactly 1 observation, 169 cells contain exactly 2 observations, etc. The 'expected' frequency distribution is computed using the Poisson approximation mentioned earlier. The observed and expected frequency distributions are quite close; the differences are what one would expect from random variation.

The 'observed' moments are computed from the observed frequency distribution. (The skewness and kurtosis have been standardized by the variance in the usual way.) These can be compared with the 'expected' moments; the true moments under the assumption of independence. These 'expected' moments are not based on the Poisson approximation; they are the exact moments. The formulas for these moments are given in the next section. To aid in comparing the observed and expected variance (1.0332 versus 0.9893); we have computed the z-score which attained the modest value of 1.004.

In summary, for this example our procedure has found no evidence of dependence/structure. This agrees with the known truth in this case.

Suppose we did not suspect *a priori* that this data set had independent columns. We would then have performed some type of transformation in Step A. Using either T1 or T2 would still lead to the conclusion that no structure is visible in this data. For example, using T2 (with the matrix **A** chosen to be upper triangular) leads to the output:

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
                 0      1      2     3    4    5    6    7    8
Observed 367.00 396.00 182.00 62.00 10.0 5.00 1.00 1.00 0.00
Expected 376.71 376.71 188.35 62.78 15.7 3.14 0.52 0.07 0.01


The moments of the distribution of cell counts are:
         mean variance skewness kurtosis
Observed    1  1.00391  1.20582  2.37976
Expected    1  0.98930  0.98396  0.94713

The z-score (conservative) for the variance of cell counts = .334
Upper bound for z-score =  1.37
```

There is a discrepancy between the observed and expected kurtosis, but this means little as the sample kurtosis is highly variable. Note that we now report two different z-scores. The transformation T2 makes the columns of **Z** (the transformed data) have sample correlations *exactly* equal to zero, which would not occur if the columns of Z were actually independent. This makes the variance of the cell counts somewhat smaller than would occur under the assumption of independence. Thus the usual z-score is conservative. A crude and heuristic degrees of freedom correction is used to obtain a larger z-score (reported in the output as the 'upper bound') which appears in simulations to be 'liberal'. For details see Section 5. In practice, the difference between these two z-scores is of little importance, for with a large sample size n, one typically does not reject the null hypothesis unless both z-scores are quite large.

As a variation on the previous example, we now consider a data set with correlated columns. The data matrix **X** is still 1024 × 10. Observations are sampled from a MVN distribution having a correlation of .2 between all pairs of covariates. If we neglect to transform this data, but apply steps B-D directly to the raw data, we obtain the following results.

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
```

```
            0      1      2      3     4    5    6    7    8
Observed 431.00 358.00 154.00 45.00 16.0 6.00 8.00 1.00 1.00
Expected 376.71 376.71 188.35 62.78 15.7 3.14 0.52 0.07 0.01

          9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Observed  0  0  1  1  0  0  0  0  0  1  0  0  0  0  0  0  1
Expected  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

The moments of the distribution of cell counts are:
         mean variance skewness kurtosis
Observed    1  2.32031  6.50023 79.12713
Expected    1  0.98930  0.98396  0.94713

The z-score for the variance of cell counts =  30.441
```

Even a rather modest correlation of .2 has dramatically altered the frequency distribution of the cell counts; a cell count of 18 or 25 is essentially impossible under independence. Thus our procedure loudly proclaims that there is structure in this data. However, the only structure in the data is the correlation structure which is removed using either transformation T1 or T2. Using T2 (with upper triangular **A**) gives us:

```
For the number of cuts = 2
The frequency distribution of the cell counts is:
            0      1      2      3     4    5    6    7    8
Observed 393.00 351.00 189.00 71.00 18.0 2.00 0.00 0.00 0.00
Expected 376.71 376.71 188.35 62.78 15.7 3.14 0.52 0.07 0.01

The moments of the distribution of cell counts are:
         mean variance skewness kurtosis
Observed    1  1.03516  0.90684  0.36108
Expected    1  0.98930  0.98396  0.94713

The z-score (conservative) for the variance of cell counts = 1.049
Upper bound for z-score =  2.101
```

This output indicates that no structure remains in the transformed data.

Examples with Randomly Located Clusters

We now consider two examples where the data consists of many randomly located clusters. We take **X** to be 1024 × 10 in both cases. The observations in **X** are made up of m clusters. The cluster centers (denoted $\mu_1, \mu_2, \ldots, \mu_m$) are independently generated from a MVN(0, **I**) distribution. The members of cluster i are independently generated from a MVN(μ_i, σ**I**) distribution. Here **I** is the 10×10 identity matrix. In both examples, the value of σ , which controls the size of the clusters, has been made large enough so that there is little evidence of clustering (or other structure) visible in the bivariate scatter plots. A careless data analyst might easily conclude there is no structure in the data.

Our analysis is given below. In both cases, the data has been transformed using T2.

The data in our first example is composed of 256 clusters, each containing 4 observations. The value of σ is 0.2 . The output for $d = 2$ is:

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
                0      1      2     3     4     5    6    7    8
Observed 538.00 225.00 105.00 80.00 48.0 19.00 3.00 4.00 2.00
Expected 376.71 376.71 188.35 62.78 15.7  3.14 0.52 0.07 0.01


The moments of the distribution of cell counts are:
         mean variance skewness kurtosis
Observed    1  1.96875  1.64175  2.67800
Expected    1  0.98930  0.98396  0.94713

The z-score (conservative) for the variance of cell counts = 22.401
Upper bound for z-score =  23.944
```

This clearly signals the existence of some type of structure in the data.

By making d larger, we can check for the existence of clustering or nonuniformity on a smaller scale. With $d = 4$, the number of cells in our contingency table is $4^{10} = 1048576$. Storing a complete contingency table this large would require too much memory. However, since $n = 1024$, the vast majority of these cells are empty. Because there is no need to keep track of the empty cells, the amount of memory we require is not excessive. Carrying out our analysis on the same data using $d = 4$ leads to:

```
For the number of cuts = 4,
The frequency distribution of the cell counts is:
               0    1    2 3
Observed 1047634  868 66.0 8
Expected 1047552 1023  0.5 0
```

```
The moments of the distribution of cell counts are:
            mean variance skewness kurtosis
Observed 0.00098  0.00115 39.47442 1855.986
Expected 0.00098  0.00098 31.99860 1023.851

The z-score (conservative) for the variance of cell counts = 128.53
Upper bound for z-score =  128.565
```

This output shows that the number of cells containing 2 or 3 observations is much larger than one would expect under independence. This concludes our discussion of the first example.

The data in our second example is composed of 16 clusters, each containing 64 observations. The value of σ is 0.7. The output for $d = 2$ is:

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
              0      1      2     3    4      5    6    7    8 9 10
Observed 473.00 301.00 134.00 61.00 30.0 14.00 4.00 3.00 1.00 1  2
Expected 376.71 376.71 188.35 62.78 15.7  3.14 0.52 0.07 0.01 0  0

The moments of the distribution of cell counts are:
         mean variance skewness kurtosis
Observed    1  1.78516  2.17656  7.29766
Expected    1  0.98930  0.98396  0.94713

The z-score (conservative) for the variance of cell counts = 18.202
Upper bound for z-score =  19.648
```

Again, our analysis clearly shows that structure exists in this data.

The weakness of the method in both examples is that it gives no clear indication of the nature of the structure which is detected.

An Example Using Speech Data

The data matrix **X** in this example is 1507×10. The data was obtained by sampling from a much larger matrix of digitized speech data consisting of 10 dimensional 'lpc' vectors. The lpc vectors in this sample all correspond to 'unvoiced' sounds.

Suppose the object of our analysis is to find out if the 'unvoiced' lpc vectors tend to lie in clusters. Examination of the bivariate scatter plots reveals some structure in the data (curvature and heteroscedasticity), but no evidence of clustering. Applying transformation T1 (principal components) to this data and then carrying out our analysis for $d = 4$ leads to the following output:

```
For the number of cuts = 4,
The frequency distribution of the cell counts is:
               0       1    2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Observed 1047111 1463.00 0.00 0 0 0 0 0 0 1  0  0  0  0  0  0  0
Expected 1047070 1504.84 1.08 0 0 0 0 0 0 0  0  0  0  0  0  0  0

         17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Observed  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
Expected  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

The moments of the distribution of cell counts are:
            mean variance  skewness    kurtosis
Observed 0.00144  0.00264 317.00704 206605.0732
Expected 0.00144  0.00144  26.37693    695.7016

The z-score (conservative) for the variance of cell counts = 611.618
Upper bound for z-score =  611.663
```

Examining the frequency distribution, we see there is one cell which contains 35 observations! There is another cell which contains 9 observations. If the coordinates were in fact independent, one would not expect to see any cells with more than 2 observations, so these two cells must be regarded as quite unusual. A little detective work reveals that these two cells are neighboring cells lying close to the center of the data set. Thus, this data set contains a small, but relatively dense cluster (containing around 44 = 35 + 9 observations) near its center. In terms of speech, I do not know what this cluster corresponds to. It may turn out to be of no importance.

Are there any other clusters in the data? In order to investigate this question, all but one (43 out of 44) of the observations in the two unusual cells were deleted from the data set. Reanalyzing the data (using transformation T1 and $d = 2$) leads to the following:

```
For the number of cuts = 2,
The frequency distribution of the cell counts is:
               0      1      2      3     4    5    6    7    8
Observed 278.00 323.00 230.00 127.00 38.00 22.0 5.00 0.00 1.00
Expected 245.13 350.46 250.52 119.39 42.67 12.2 2.91 0.59 0.11

            9
Observed 0.00
Expected 0.02
```

```
The moments of the distribution of cell counts are:
            mean variance skewness kurtosis
Observed 1.42969  1.65521  0.96481  1.00909
Expected 1.42969  1.41437  0.82291  0.66243

The z-score (conservative) for the variance of cell counts = 3.847
Upper bound for z-score =  4.962
```

This suggests that some structure exists (which we already know from examining the bivariate scatter plots), but that there is no clustering which is as pronounced or dramatic as that found in the earlier simulated examples. If any clusters remain in this data set, they are not very well defined.

4 The Null Distribution

In this section we shall give expressions for the moments and distribution of U_π, the number of observations contained in cell π. We also present a formula for the variance of the quantity W defined in equation 2.3. All these results are derived under the assumption that the coordinates are independent; under this assumption the results are exact. If a data-dependent transformation (such as T1 or T2) has been applied to the data, the results can only be regarded as approximations. Proofs for these results are given in an appendix.

The following notation will be useful. For any real number x and positive integer k we define

$$\langle x \rangle_k = \prod_{j=0}^{k-1} (x - j) \tag{4.1}$$

so that, for instance, the combinatorial coefficient $\binom{x}{k}$ may be written as $\langle x \rangle_k / k!$.

Let ξ_k be the probability that the first k observations belong to the cell π. More formally,

$$\xi_k \equiv \Pr\{t_{i\cdot} = \pi \text{ for } 1 \leq i \leq k\}$$

where $t_{i\cdot}$ denotes the i^{th} row of $\mathbf{T}$. Using urn model arguments (sampling without replacement) it is easy to show that

$$\xi_k = \left(\frac{\langle n/d \rangle_k}{\langle n \rangle_k} \right)^p .$$

The factorial moments of U_π can now be given as

$$E\langle U_\pi \rangle_k = \langle n \rangle_k \, \xi_k . \tag{4.2}$$

The ordinary moments (about the origin) can be obtained directly from the factorial moments. The general formula is given in the appendix. As special cases we note that

$$\begin{aligned}
EU_\pi &= n\xi_1 = \frac{n}{d^p} \\
EU_\pi^2 &= n\xi_1 + n(n-1)\xi_2 \\
EU_\pi^3 &= n\xi_1 + 3n(n-1)\xi_2 + n(n-1)(n-2)\xi_3 \\
EU_\pi^4 &= n\xi_1 + 7n(n-1)\xi_2 + 6n(n-1)(n-2)\xi_3 + n(n-1)(n-2)(n-3)\xi_4
\end{aligned}$$

Finally, the central moments may be obtained from the moments about the origin via the formula

$$E(U_\pi - \mu)^k = \sum_{j=0}^{k} \binom{k}{j} (-\mu)^j EU_\pi^{k-j}$$

This is the route used to calculate the moments reported in the computer output in the Examples section. The skewness and kurtosis are, as usual, defined by

$$\text{skewness} = \frac{E(U_\pi - \mu)^3}{\sigma^3}$$

and

$$\text{kurtosis} = \frac{E(U_\pi - \mu)^4}{\sigma^4} - 3$$

where μ and σ^2 are the mean and variance of U_π.

The exact distribution of the cell counts is given by

$$\Pr\{U_\pi = k\} = \binom{n}{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j \xi_{k+j} \qquad (4.3)$$

where $\xi_0 \equiv 1$. In practice, the series is truncated when the terms become sufficiently small. This series should give accurate results when the terms decay to zero rapidly, but the alternating signs may render the formula useless when the rate of decay is slow. The expected frequency distribution reported in the computer output in the Examples section was obtained from the Poisson approximation in (2.2) .

We now show how to compute the quantities μ_w and σ_w^2 occurring in equation 2.4 for the z-scores. We shall freely use the ordinary and factorial moments which are easily calculated using the previously given formulas. First the mean:

$$\mu_w \equiv EW = EU_\pi^2 - \left(\frac{n}{d^p}\right)^2 \qquad (4.4)$$

Now for the variance. Define

$$Q_{jk} = \langle n \rangle_{j+k} \left\{ \left(\frac{1}{d}\right) \frac{\langle n/d \rangle_{j+k}}{\langle n \rangle_{j+k}} + \left(1 - \frac{1}{d}\right) \frac{\langle n/d \rangle_j \langle n/d \rangle_k}{\langle n \rangle_{j+k}} \right\}^p .$$

Then we can write

$$\begin{aligned}
\sigma_w^2 \equiv \operatorname{Var}(W) &= Q_{22} + 2Q_{12} + Q_{11} - (EU_\pi^2)^2 \\
&\quad + \frac{1}{d^p} \left(4E\langle U_\pi \rangle_3 + 6E\langle U_\pi \rangle_2 + EU_\pi\right) .
\end{aligned} \qquad (4.5)$$

This expression is easy to compute with, but rather difficult to comprehend. When n and d^p are both large, useful approximations are given by

$$\mu_w \approx \frac{n}{d^p} \text{ and } \sigma_w^2 \approx \frac{2n^2}{d^{3p}} . \tag{4.6}$$

Note that $W = (n/d^{2p})\chi^2$ where χ^2 is the usual test statistic for independence in contingency tables. Thus (4.6) corresponds to the fact that the variance of a χ^2 random variable is twice its mean. These formulas are based upon d^p degrees of freedom and will somewhat overstate the actual mean and variance.

5 The Upper Bound for z-scores

As noted earlier, if the data has been transformed (perhaps by using T1 or T2) to remove all correlations, the z-scores computed from (2.4) will be conservative (too small). The following correction tends to produce a z-score which is liberal (too large). The two z-scores give us a likely range of values which can be used to carry out crude hypothesis tests.

Define

$$r = 1 - \left[\frac{\binom{p}{2}}{d^p - p(d-1) - 1} \right] .$$

The liberal z-score is given by

$$z = \frac{W - r\mu_w}{\sigma_w \sqrt{r}} . \tag{5.1}$$

The values of μ_w and σ_w are those defined in the previous section. Note that $\binom{p}{2}$ is the number of correlations which are estimated, that is, removed from the data, and $d^p - p(d-1) - 1$ is the number of degrees of freedom in our $d \times d \times \cdots \times d$ contingency table.

APPENDIX

In section 2 (step B), we defined the $n \times p$ matrix $\mathbf{T}$. Under the assumption that the coordinates are independent, it is clear that the columns of $\mathbf{T}$ are independent. In the process of constructing $\mathbf{T}$, we assume that d divides n exactly. Thus each column consists of values $\{1, 2, \ldots, d\}$ with n/d repetitions of each value. and entries in a column are exchangeable.

Using urn model arguments (sampling without replacement) and by independence of columns, we have

$$\xi_k = \left[\frac{\langle n/d \rangle_k}{\langle n \rangle_k} \right]^p . \qquad (A.1)$$

Let C_k^n be the class of subsets in $\{1, 2, \ldots, n\}$ with the cardinality k, i.e.

$$C_k^n = \{\sigma \subset \{1, 2, \ldots. n\} : |\sigma| = k\}$$

for $k = 1, 2, \ldots, n$, where $|\sigma|$, the cardinality of σ, is the number of elements in the set σ. Then it is easy to show the relation

$$\binom{U_\pi}{k} = \sum_{\sigma \in C_k^n} I\,(\,t_{i.} = \pi, \forall\, i \in \sigma), \qquad (A.2)$$

where $t_{i.}$ denotes the i^{th} row of $\mathbf{T}$. By exchangeability of rows, we have

$$E\binom{U_\pi}{k} = \binom{n}{k} \xi_k . \qquad (A.3)$$

From $(A.3)$, the factorial moments of U_π becomes

$$E\langle U_\pi \rangle_k = \langle n \rangle_k \xi_k . \qquad (A.4)$$

The ordinary moments (about the origin) can be obtained directly from the factorial moments. This yields

$$EU_\pi^k = \sum_{j=1}^{k} S_k^{(j)} \langle n \rangle_j \, \xi_j \, , \qquad (A.5)$$

where $S_k^{(j)}$ is a Stirling number of the second kind ; see Abramowitz and Stegun (1970) for definitions and tables of these numbers.

To derive the exact distribution of U_π, we can use the relation

$$I\,(\,U_\pi = k) = \sum_{\sigma \in C_k^n} I\,(t_{i.} = \pi, \forall\, i \in \sigma,\; t_{i.} \neq \pi, \forall\, i \notin \sigma).$$

Thus by exchangeability of rows, we have

$$\begin{aligned} P(U_\pi = k) &= \binom{n}{k} E\left[\prod_{i=1}^{k} I(t_{i.} = \pi) \prod_{i=k+1}^{n} I(t_{i.} \neq \pi)\right] \\ &= \binom{n}{k} E\left[\prod_{i=1}^{k} I(t_{i.} = \pi) \prod_{i=k+1}^{n} \{1 - I(t_{i.} = \pi)\}\right] \\ &= \binom{n}{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j \xi_{k+j}. \qquad (A.6) \end{aligned}$$

We now show how to derive the variance of W in (2.3). First, we will simplify the formula for $Var(W)$.

$$\begin{aligned} Var(W) &= Var\left(\frac{1}{d^p}\sum_\pi U_\pi^2 - \left(\frac{n}{d^p}\right)^2\right) \\ &= Var\left(\frac{1}{d^p}\sum_\pi U_\pi^2\right) \\ &= E\left\{\frac{1}{d^{2p}}\left(\sum_\pi U_\pi^2\right)^2\right\} - \left\{E\left(\frac{1}{d^p}\sum_\pi U_\pi^2\right)\right\}^2 \\ &= E\left(U_{\Pi_1}^2, U_{\Pi_2}^2\right) - \left(EU_\pi^2\right)^2, \qquad (A.7) \end{aligned}$$

where Π_1, Π_2 are iid uniform random variables on the set of all possible cells. Now we can derive the relation

$$\begin{aligned} &\sum I(t_{i.} = \Pi_1, \forall\, i \in \sigma_1,\, t_{i.} = \Pi_2, \forall\, i \in \sigma_2) \\ &= \binom{U_{\Pi_1}}{j}\binom{U_{\Pi_2}}{k} + \left[\binom{U_{\Pi_1}}{j,k} - \binom{U_{\Pi_1}}{j}\binom{U_{\Pi_2}}{k}\right]\delta_{\Pi_1,\Pi_2}, \qquad (A.8) \end{aligned}$$

where $\delta_{\Pi_1\Pi_2}$ is the Kronecker delta, and the summation on the left hand side is over all $\sigma_1 \in C_j^n, \sigma_2 \in C_k^n$ with $\sigma_1 \cap \sigma_2 = \emptyset$. First, take expectation on the left hand side of equation $(A.8)$. Then by exchangeability of rows, it becomes

$$\binom{n}{j,k} P(t_{i.} = \Pi_1,\, 1 \le i \le j,\, t_{i.} = \Pi_2,\, j+1 \le i \le j+k). \qquad (A.9)$$

Using the independence of columns of $\mathbf{T}$, and using the fact that if Π_1 and Π_2 are iid uniform random variables, then coordinates of Π_1 are independent of those of Π_2, we can show that $(A.9)$ becomes

$$\binom{n}{j,k}\left\{\frac{1}{d}\frac{\langle n/d\rangle_{j+k}}{\langle n\rangle_{j+k}} + \left(1 - \frac{1}{d}\right)\frac{\langle n/d\rangle_j \langle n/d\rangle_k}{\langle n\rangle_{j+k}}\right\}^p. \qquad (A.10)$$

Now take expectation on the right hand side of equation $(A.8)$ to get

$$E\binom{U_{\Pi_1}}{j}\binom{U_{\Pi_2}}{k} + \frac{1}{d^p}E\left[\binom{U_\pi}{j,k} - \binom{U_\pi}{j}\binom{U_\pi}{k}\right]. \qquad (A.11)$$

Define

$$Q_{jk} = \langle n\rangle_{j+k}\left\{\frac{1}{d}\frac{\langle n/d\rangle_{j+k}}{\langle n\rangle_{j+k}} + \left(1-\frac{1}{d}\right)\frac{\langle n/d\rangle_j\langle n/d\rangle_k}{\langle n\rangle_{j+k}}\right\}^p.$$

By combining $(A.10)$ and $(A.11)$ and then by multiplying $j!k!$ on both sides, we have

$$E\{\langle U_{\Pi_1}\rangle_j\langle U_{\Pi_2}\rangle_k\} + \frac{1}{d^p}E\{\langle U_\pi\rangle_{j+k} - \langle U_\pi\rangle_j\langle U_\pi\rangle_k\} = Q_{jk}$$

or

$$E\{\langle U_{\Pi_1}\rangle_j\langle U_{\Pi_2}\rangle_k\} = Q_{jk} + \frac{1}{d^p}E\{\langle U_\pi\rangle_j\langle U_\pi\rangle_k - \langle U_\pi\rangle_{j+k}\}. \qquad (A.12)$$

Thus

$$\begin{aligned}
Var(W) &= E\left(U^2_{\Pi_1}U^2_{\Pi_2}\right) - \left(EU^2_\pi\right)^2\\
&= E\{\langle U_{\Pi_1}\rangle_2\langle U_{\Pi_2}\rangle_2 + 2U_{\Pi_1}\langle U_{\Pi_2}\rangle_2 + U_{\Pi_1}U_{\Pi_2}\} - \left(EU^2_\pi\right)^2\\
&= Q_{22} + 2Q_{12} + Q_{11} - \left(EU^2_\pi\right)^2\\
&+ \frac{1}{d^p}E\{\langle U_\pi\rangle_2\langle U_\pi\rangle_2 + 2U_\pi\langle U_\pi\rangle_2 + U^2_\pi - \langle U_\pi\rangle_4 - 2\langle U_\pi\rangle_3 - \langle U_\pi\rangle_2\}\\
&= Q_{22} + 2Q_{12} + Q_{11} - \left(EU^2_\pi\right)^2 + \frac{1}{d^p}E\{U^4_\pi - \langle U_\pi\rangle_4 - 2\langle U_\pi\rangle_3 - \langle U_\pi\rangle_2\}\\
&= Q_{22} + 2Q_{12} + Q_{11} - \left(EU^2_\pi\right)^2 + \frac{1}{d^p}\{4E\langle U_\pi\rangle_3 + 6E\langle U_\pi\rangle_2 + EU_\pi\} \qquad (A.13)
\end{aligned}$$

The statistic W is proportional to the usual χ^2 test statistic for independence in contingency tables. When $p = 2$, our formula for $Var(W)$ is equivalent to a special case of the Haldane-Dawson formula (Haldane. 1939; Dawson, 1954) for the variance of χ^2.

In order to calculate an approximate value of $Var(W)$ for large n and d^p, we can use the following approximation : for large y and for small k and j,

$$\frac{\langle y\rangle_{j+k}}{\langle y\rangle_j\langle y\rangle_k} = \frac{\langle y-j\rangle_k}{\langle y\rangle_k} = \prod_{i=0}^{k-1}\left(1-\frac{j}{y-i}\right) \approx 1 - \sum_{i=0}^{k-1}\frac{j}{y-i} \approx 1 - \frac{kj}{y}. \qquad (A.14)$$

Thus

$$\begin{aligned}
Q_{jk} &= \langle n\rangle_{j+k}\left\{\frac{1}{d}\frac{\langle n/d\rangle_{j+k}}{\langle n\rangle_{j+k}} + \left(1-\frac{1}{d}\right)\frac{\langle n/d\rangle_j\langle n/d\rangle_k}{\langle n\rangle_{j+k}}\right\}^p\\
&= \langle n\rangle_j\langle n\rangle_k\left\{\frac{\langle n/d\rangle_j\langle n/d\rangle_k}{\langle n\rangle_j\langle n\rangle_k}\right\}^p\frac{\langle n\rangle_{j+k}}{\langle n\rangle_j\langle n\rangle_k}\left\{\frac{\langle n\rangle_j\langle n\rangle_k}{\langle n\rangle_{j+k}}\right\}^p\left\{\frac{1}{d}\frac{\langle n/d\rangle_{j+k}}{\langle n/d\rangle_j\langle n/d\rangle_k} + 1 - \frac{1}{d}\right\}^p\\
&\approx E\langle U_\pi\rangle_j E\langle U_\pi\rangle_k\left\{1-\frac{kj}{n}\right\}\left\{1-\frac{kj}{n}\right\}^{-p}\left\{1-\frac{kj}{n}\right\}^p\\
&= \left(1-\frac{kj}{n}\right)E\langle U_\pi\rangle_j E\langle U_\pi\rangle_k. \qquad (A.15)
\end{aligned}$$

Hence

$$\begin{aligned}
&Q_{22} + 2Q_{12} + Q_{11} - \left(EU_\pi^2\right)^2 \\
&= Q_{22} - \left\{E\langle U_\pi\rangle_2\right\}^2 + 2\left\{Q_{12} - EU_\pi E\langle U_\pi\rangle_2\right\} + Q_{11} - \left(EU_\pi\right)^2 \\
&\approx -\frac{1}{n}\left\{2E\langle U_\pi\rangle_2 + EU_\pi\right\}^2 . \qquad (A.16)
\end{aligned}$$

Finally by using $E\langle U_\pi\rangle_k \approx n^k/d^{kp}$, we have

$$\begin{aligned}
Var(W) &\approx -\frac{1}{n}\left\{2E\langle U_\pi\rangle_2 + EU_\pi\right\}^2 + \frac{1}{d^p}\left\{4E\langle U_\pi\rangle_3 + 6E\langle U_\pi\rangle_2 + EU_\pi\right\} \\
&\approx -\frac{1}{n}\left\{2\frac{n^2}{d^{2p}} + \frac{n}{d^p}\right\}^2 + \frac{1}{d^p}\left\{4\frac{n^3}{d^{3p}} + 6\frac{n^2}{d^{2p}} + \frac{n}{d^p}\right\} \\
&= 2\frac{n^2}{d^{3p}} . \qquad (A.17)
\end{aligned}$$

References

Abramowitz, M. and Stegun, I.A. (1970) *Handbook of Mathematical Functions*. Dover, New York.

Dawson, R.B. (1954) A simplified expression for the variance of the χ^2-function on a contingency table. *Biometrika* **41**, 280.

Friedman, J.H. (1987) Exploratory Projection Pursuit. *Journal of the American Statistical Association* **82**, 249-266.

Haldane, J.B.S. (1939) The mean and variance of χ^2 when used as a test of homogeneity when expectations are small. *Biometrika* **31**, 346-65.

Huber, P.J. (1985) Projection Pursuit. *Annals of Statistics* **13**, 435-475.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
447

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A Preliminary Test For Structure In Large, High-Dimensional Data Sets

5. TYPE OF REPORT & PERIOD COVERED
TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Fred W. Huffer and Cheolyong Park

8. CONTRACT OR GRANT NUMBER(s)
N00014-89-J-1627

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-042-267

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Statistics & Probability Program Code 1111

12. REPORT DATE
September 5, 1991

13. NUMBER OF PAGES
20

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Exploratory data analysis, testing for independence, detecting structure, multivariate analysis.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We present a natural preliminary test for the presence of structure (nontrivial dependence) in a data set, and give some examples of its use. The procedure consists of sphering the data to remove correlations, then binning or discretizing the data, and finally, studying the cell counts in the resulting contingency table. If this procedure detects structure, we can then use more computationally intensive methods to determine the nature of this structure.